DOCUMENT RESUME

ED 389 737 TM 024 366

AUTHOR Donoghue, John R.; Mazzeo, John
TITLE Assessing Some of the Properties of Longer Blocks in the 1992 NAEP Reading Assessment.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
REPORT NO ETS-RR-95-28
PUB DATE Aug 95
NOTE 27p.; Version of a paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Chi Square; *Goodness of Fit; Grade 8; Grade 12; National Surveys; Reading Ability; *Reading Tests; Robustness (Statistics); *Scaling; Secondary Education; *Structural Equation Models; *Test Construction; Test Format
IDENTIFIERS *Blocks; *National Assessment of Educational Progress

ABSTRACT

At grades 8 and 12, the 1992 National Assessment of Educational Progress (NAEP) reading assessment contained a small number of 50-minute blocks in addition to the usual 25-minute blocks. To determine whether to incorporate the 50-minute blocks into the operational scaling, this study sought to determine whether the longer blocks measured a different construct from that assessed by the 25-minute blocks. Structural equation modeling tested the hypothesis that the structural parameters relating reading ability to demographic variables do not differ across block type. A multiple group analysis, where type of block (25-minute or 50-minute) defined the two groups, was used. The null hypothesis was that the two types of blocks measure the same trait but could differ in observed mean and variance. Results of the main analysis did not reject the hypothesis of invariant structural parameters, and so the 50-minute blocks were not incorporated into the 1992 NAEP scales. Sensitivity analyses indicated that this conclusion was moderately robust to assumptions made about missing data for items that were not reached. Analyses using other measures of fit yielded similar results, although the magnitude of chi-square statistics was affected by the fit measure chosen. (Contains 6 tables and 13 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 389 737

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

ASSESSING SOME OF THE PROPERTIES OF LONGER BLOCKS IN THE 1992 NAEP READING ASSESSMENT

John R. Donoghue
John Mazzeo

BEST COPY AVAILABLE

TM024306

Educational Testing Service
Princeton, New Jersey
August 1995

Assessing Some of the Properties of Longer Blocks in the 1992 NAEP Reading Assessment

John R. Donoghue and John Mazzeo

Educational Testing Service

The work upon which this document is based was performed for the National Center for Education Statistics, Office of Educational Research and Improvement, by Educational Testing Service. An earlier version of this paper was presented at the symposium "Analysis of Innovative Data Structures in the National Assessment of Educational Progress," at the annual meeting of the American Educational Research Association in Atlanta, GA, April 1993. The authors would like to thank Don Rock and Howard Wainer for their helpful comments.

Copyright © 1995. Educational Testing Service. All rights reserved.

Abstract

At Grades 8 and 12, the 1992 NAEP reading assessment contained a small number of 50 minute blocks, in addition to the usual 25 minute blocks. In order to determine whether to incorporate the 50 minute blocks into the operational scaling, this study sought to determine whether longer blocks measure a different construct from that assessed by the 25 minute blocks. Structural equation modeling tested the hypothesis that the structural parameters relating reading ability to demographic variables do not differ across block type. A multiple group analysis, where type of block (25 minute or 50 minute) defined the two groups, was used. The null hypothesis for this comparison was that the two types of blocks measure the same trait but may differ in observed mean and variance.

Results of the main analysis did not reject the hypothesis of invariant structural parameters, and so the 50 minute blocks were incorporated in the 1992 NAEP scales. Sensitivity analyses indicated that this conclusion was moderately robust to assumptions made about missing data for items which were not reached, although Grade 8 results were more robust than those for Grade 12. Analyses using other measures of fit yielded the same pattern of results, although the magnitude of the χ^2 statistics were affected by the fit measure chosen, particularly the asymptotically distribution free method.

Attempts to replicate the main analysis in independent samples yielded similar χ^2 values at Grade 8, but Grade 12 yielded χ^2 values which were substantially higher for some of the samples. The Grade 12 results raise questions as to the generalizability of the main analysis. Alternatives to the reliance on χ^2 measures are discussed for future research.

Overview

The National Assessment of Educational Progress (NAEP) is a federally mandated survey of what American students at Grades 4, 8, and 12 know and can do. The NAEP contract is conducted by Educational Testing Service under the direction of the National Assessment Governing Board, and administered by the National Center of Education Statistics. The 1992 NAEP reading assessment is based on a new set of objectives and specifications, developed by a consensus process (NAGB, 1991). In contrast to previous, unidimensional scales, the 1992 NAEP reading framework calls for three subscales based on purposes of reading: reading for literary experience, reading to be informed, and reading to perform a task.

Compared to earlier NAEP reading assessments, the 1992 assessment also contains longer reading passages which are intended to be more authentic examples of reading tasks encountered in and out of school. In addition to multiple choice items, each passage is followed by a number of constructed response items, accounting for over one-half of the assessment time. Some of these items are relatively short constructed response items, requiring a sentence or a paragraph response. These short constructed response items are typically scored as correct or incorrect. In addition, each reading passage contains at least one extended constructed response item, which requires a more in-depth, elaborated response. These extended constructed response items were score polytomously:

0 - Unsatisfactory;

1 - Partial;

2 - Essential;

3 - Extensive, which demonstrates more in-depth understanding.

Detailed scoring rubrics were developed for each polytomous item. The actual items are secure, and so cannot be reproduced here. However, a typical extended constructed response item might ask the

examinee to compare and contrast two accounts of a historical event, or to describe the feelings of a character in a story and describe the events in the story which triggered those feelings.

NAEP uses a balanced incomplete block (BIB) design. Separately timed sections, termed blocks, are combined to form booklets according to the BIB design. The individual booklets are spiraled, i.e., assigned to examinees according to a systematic arrangement such that each booklet is presented to a randomly equivalent group of examinees (see Messick, Beaton, & Lord, 1983, for more details). To assess the proficiency of a population and important subgroups, BIB spiraling is very efficient; it allows a large number of items to be presented, while simultaneously limiting the testing time for an individual examinee. However, relatively little information is obtained for individual examinees. NAEP uses item response theory (IRT) to pull together the pieces of the BIB spiral assessment, to establish vertical (cross-grade) scales, and to perform trend analyses.

The Problem

The majority of the 1992 NAEP reading assessment consists of separately timed blocks of cognitive items, requiring 25 minutes each. Individual students were administered two such cognitive blocks, in addition to a number of demographic items, and questions concerning their educational background and the educational practices to which the student had been exposed.

At Grades 8 and 12, the "reading to be informed" subscale also contained a few extended blocks, requiring 50 minutes of administration time each. While the 25 minute blocks were based on a single reading passage, a typical 50 minute block presented students with two passages. These 50 minute blocks reflect reading specialists' desire to incorporate longer, more realistic passages, which allow the examinees to interact with the material in more depth. They are intended to reflect more accurately the type of reading tasks regularly encountered by students in and out of school.

Because the 1992 reading is a new assessment, we were particularly interested in closely

BEST COPY AVAILABLE

examining its properties. We would like to incorporate the 50 minute blocks into the operational scaling. However, it should be demonstrated that the longer blocks do not measure a different construct from that assessed by the 25 minute blocks.

The Comparison

Unfortunately, the reading design is such that no examinee takes both 25 and 50 minute blocks, making a direct verification impossible. Nor are there common items across block types. This excludes IRT- and factor analysis-based measures, such as a likelihood ratio test of the fit of one trait versus the fit with separate traits for the two block types.

However, the NAEP sampling design does insure that randomly equivalent groups take the 25 and 50 minute blocks. This allows an alternative way to examine the question of identity of constructs. If the 50 minute blocks measure the same construct as the 25 minute blocks, then, after adjusting for differences in reliability, they should exhibit identical relationships with important demographic variables, such as gender, race/ethnicity, and parents' educational status.

Structural equation modeling can be used to test the hypothesis that the structural parameters relating reading ability to demographic variables are identical across block type. A multiple group analysis, where type of block (25 minute or 50 minute) defined the two groups, was used. The null hypothesis for this comparison is that the two types of blocks constitute congeneric measures; they measure the same trait but may differ in observed mean and variance. To increase the comparability, the comparison of block types only involved the 25 minute blocks with the same reading purpose; reading to be informed.

Method

We used a two group confirmatory factor analysis model to test the hypothesis that the two types of blocks (25 minute and 50 minute) have the same relationships with background variables. The main analysis (upon which the operational decisions were based) will be described first. Then a set of additional analyses will be described.

Data

The data source for this study was a subset of the 1992 National NAEP reading assessment. Descriptive information about the blocks used in these analyses is given in Table 1. More detailed information is presented below.

Insert Table 1 about here

At each grade, two of the booklets of the BIB design were selected. These booklets consisted of only those blocks which assessed the same subscale (reading to be informed) as did the 50 minute blocks. To control for possible warmup/fatigue effects, only the first 25 minute block in each booklet was used. Two separate blocks, designated 25_A and 25_B were used. Note that blocks with the same designation were not necessarily the same blocks for the two grades.

Each of the 50 minute blocks was contained in a single booklet which required the full assessment time; no other cognitive items were administered. These booklets were administered to samples of examinees who were randomly equivalent to those who took the 25 minute blocks. Approximately four times as many examinees received a given 50 minute block as received one of the selected booklets containing Block 25_A or Block 25_B. To maintain comparability with the benchmark results comparing Block 25_A to Block 25_B (see below), a one-quarter systematic sample was drawn for each 50 minute block

by selecting every fourth examinee in the data file. The remaining 75% of the examinees for each 50 minute booklet were used to form the three replicate samples (see below).

For the main analysis, all omitted and not reached items were treated as incorrect. This scoring method is often used in item analysis and classroom testing. By treating all not reached items as incorrect, this method functions as if speed and power are perfectly negatively correlated. Missing data for background items were treated by listwise deletion. Fewer than 1.5% of the examinees were deleted from any analysis.

Background variables were selected to have a high zero-order correlation with reading ability, but not correlate too highly with other background variables (i.e., have high incremental R^2). Table 2 gives the correlations of background variables with reading ability, defined as total score on the block of cognitive reading items. The same background variables were used for both Grade 8 and Grade 12.

Insert Table 2 about here

Analyses

Four testlets were formed from the cognitive items in each block. Three of the testlets were defined as the sum of 3-5 dichotomous (multiple choice and short constructed response) items. To the degree possible, the composition of the testlets was balanced as to item type and order within the block. The fourth testlet in each block was defined as the sum of all polytomous (extended constructed response) items. Blocks contained from 1-3 polytomous items, each of which was scored on a four point scale.

Most variables, including the testlets, were treated as ordinal indicators of an underlying, normally distributed latent variable, and tetrachoric and polychoric correlations were computed. However, the ordinal formulation did not make sense for gender and the race/ethnicity indicator variables. Thus, Pearson correlations were computed among these variables, and biserial and polyserial correlations were

computed with the other, ordinal variables. Correlations were computed using PRELIS 1.1 (Joreskog & Sorbom, 1988).

The four testlets were modeled as indicators of reading ability, and all loaded on a single reading factor. Background variables were treated as if they were measured without error, and each loaded 1.0 on a separate "latent" variable. The "correlations free" model allowed the correlations of the eight background variables with reading ability to differ across block types, while the test model constrained these eight correlations to be equal across the two block types.

In addition to the model comparing a 25 minute block to a 50 minute block, another set of models was fit comparing Block 25_A to Block 25_B at each grade. These models allowed us to examine the behavior of the test statistic when the null hypothesis was true, and so provided us with a benchmark result.

LISREL 7 (Joreskog & Sorbom, 1989) was used to fit all models in this study. For the main analyses, the generalized least squares (GLS) method was used to fit models. Hu, Bentler, and Kano (1992) report that GLS was more robust to non-normality than was ML, while elliptical least squares and asymptotically distribution free methods both required larger samples than available here for the χ^2 statistic to function properly. Thus, we used GLS as the best of the readily available methods.

For the comparisons below, the χ^2 and its probability are those produced by the LISREL program. It should be borne in mind that the χ^2-statistics produced by LISREL assume simple random sampling of observations. However, NAEP uses a complex, multi-stage sampling design. This results in dependence among the observations, and so the reported χ^2-statistics are too large (e.g., Rao & Thomas, 1991), and their associated probabilities are too small. Thus, the significance tests for comparing models are liberal, and will tend to reject the null hypothesis too often.

Additional Analyses

After the main analyses were completed, and the operational decisions regarding scoring had been made, a number of subsequent analyses were conducted to further describe the 25 and 50 minute blocks. These analyses may be divided into two types, sensitivity analyses and independent replications of the 25/50 minute block comparison.

Sensitivity A number of additional analyses were conducted to assess the sensitivity of the results to specific decisions that were made in the main analysis. We chose to focus on two aspects of the main analysis, the treatment of missing data, and the choice of fit statistic.

To assess the sensitivity to the assumptions made about the missing data, a second version of each data set was constructed. In this second version, each not reached response was imputed with probability equal to the overall probability for that item (probability of a correct response for dichotomous items, and multinomial probability for polytomous items). Omitted items were still treated as incorrect, with the exception of multiple choice items, which were imputed correct with probability .25 (1 over the number of alternatives). This approach treats not reached items as missing completely at random; i.e., it treats speed and power as independent. Thus, this analysis complements the main analysis. In one sense, the two analyses are the extremes of a continuum of reasonable assumptions which might be made about not reached data.

The second focus of sensitivity was the fit function, and its associated χ^2 statistic. In order to assess this, each model was also fit using maximum likelihood (ML) and asymptotic distribution free (ADF) methods, and the χ^2 values were examined to determine how sensitive our conclusions were to the differences in methods of fit.

Replication The second set of analyses sought to assess the stability of the findings; how sample dependent are the results? For each 50 minute block, we constructed three additional, replicate data sets

BEST COPY AVAILABLE

from the 75% of the data that were left over from the main analysis. Each of the replicate data sets was then compared to Block 25_A and Block 25_B for that grade, using the same methods as in the main analysis.

Results

To simplify the presentation, the order of results will follow the chronological order in which the analyses were done. We will present the results for each of the 50 minute blocks separately. The main analysis for each block will be presented, along with the decision that was made concerning that block. This will be followed by the results of the sensitivity analysis, and finally the independent replications.

Grade 8

The top portion of Table 3 presents the results of fitting the models to the Grade 8 data. The difference test is not significant for comparing the 50 minute block to Block 25_A ($\chi^2(8) = 10.63$, $p > .05$), nor was the comparison to Block 25_B significant ($\chi^2(8) = 12.37$, $p > .05$). The null model comparing Block 25_A with Block 25_B yielded a difference statistic which was similar ($\chi^2(8) = 11.40$, $p > .05$) to the two test values. Based on these results, we concluded that the 25 and 50 minute blocks yielded similar relationships of reading with background variables, and so this block was included in the operational NAEP analysis.

Insert Table 3 about here

The middle section of Table 3 gives the fit statistics χ^2 for each of the sensitivity analyses. The absolute magnitude of the χ^2 statistics differs for the various analyses; the values are similar for ML and somewhat smaller for the imputed data. Surprisingly, while the χ^2 values for ADF are much smaller than the main analysis, the difference test is much larger. However, in each case the null comparison of 25_A with 25_B is similar to or larger than comparison with the 50 minute block. The results are fairly robust to

these assumptions; the pattern of values of the difference test leads us to the same conclusions for each analysis.

The bottom section of Table 3 gives the fit values for analysis of each of the three replicate samples of the 50 minute block. Although one of the tests is marginally significant, in general the values for each of the replicates are similar to those for the main analysis, adding further support for the operational decisions.

Grade 12, Block 50_A

Table 4 summarizes the results of analysis of Block 50_A. The top portion presents the results of fitting the test models to these data. The difference test is not significant for comparing Block 50_A to Block 25_A ($\chi^2(8) = 14.76$, $p > .05$). However, the comparison of Block 50_A to Block 25_B was marginally significant ($\chi^2(8) = 15.76$, $p = .047$). The null model comparing Block 25_A with Block 25_B yielded a difference statistic which was somewhat smaller ($\chi^2(8) = 7.92$, $p \approx .5$), but it was similar to the two test values. Although the fit was not as good as that found for the 50 minute block at Grade 8, there was not sufficient evidence to conclude that Block 50_A has different relationships with the background variables. Based on these results, we decided to include Block 50_A in the operational NAEP scaling. However, we also noted that there is more evidence of difference than there was at Grade 8.

Insert Table 4 about here

The middle section of Table 4 gives the fit statistics for each of the sensitivity analyses. ML gives similar results to those of the main analysis. The pattern of results for ADF parallels those found for Grade 8. Again, the absolute magnitude of the χ^2 statistics is smaller, but the value of the difference test is larger. In this case, however, ADF would lead us to reject the hypothesis that the two block types have the same relationships with the background variables. Thus, our conclusions are not completely robust to

method of estimation. Finally, the alternative assumption about missing data leads to a similar pattern of results as did the main analysis; comparison with one of the 25 minute blocks yields a significant χ^2, while comparison with the other is not significant, and the null comparison of Block 25_A with Block 25_B is not significant. However, the test statistic ($\chi^2(8) = 21.93$, $p < .01$) is large enough that this analysis might have led us to a different decision than was made for the main analysis. Thus, the analysis of Block 50_A is not completely robust to assumptions about missing data.

The bottom section of Table 4 gives the χ^2 fit values for analysis of each of the three replicate samples of Block 50_A. The χ^2 difference values for Replicate A and for the comparison of Replicate B with Block 25_A are much larger than those for the main analysis, while the remaining values are similar to or smaller than those of the main analysis. If our decision had been based on analysis of Replicate A, we would have rejected the hypothesis that Block 50_A has the same relationship with background variables. On the other hand, the opposite conclusion would be reached from both the main sample and Replication C, while Replicate B would have left us unsure. This troubling difference will be discussed in more detail below.

Grade 12, Block 50_B

The top portion of Table 5 presents the results of fitting the models to assess Block 50_B of the Grade 12 data. The difference test was not significant for comparing Block 50_B block to Block 25_A ($\chi^2(8) = 7.34$, $p > .05$), nor was the comparison to Block 25_B significant ($\chi^2(8) = 11.78$, $p > .05$). The null model comparing Block 25_A with Block 25_B yielded a difference statistic which was similar ($\chi^2(8) = 7.92$, $p > .05$) to the two test values. Based on these results, we concluded that Block 50_B and the 25 minute blocks yielded similar relationships of reading with background variables, and so it was included in the operational analysis of the 1992 NAEP reading assessment.

Insert Table 5 about here

The middle section of Table 5 gives the χ^2 fit statistics for each of the sensitivity analyses. As was the case for the other two 50 minute blocks, the absolute magnitude of the χ^2 statistics differs for the various analyses, and the value of the difference test is somewhat higher for the imputed values and ADF. However, only the ADF comparison with Block 25_B reaches statistical significance. The results are fairly robust to these assumptions; we would reach similar conclusions for each of the analyses.

The bottom section of Table 5 gives the χ^2 fit values for analysis of each of the three replicate samples of Block 50_B. The χ^2 difference values for Replicates A and B are much larger than those for the main analysis, while those for Replicate C are noticeably smaller. If our decision had been based on analysis of Replicates A or B, we would have rejected the hypothesis that Block 50_B has the same relationship with background variables. On the other hand, the opposite conclusion would be reached from both the main sample and Replicate C. As was the case with Block 50_A, this difference is troubling, and indicates a weakness in decisions based solely on the χ^2 test.

Analyses of Full Data

Due to the variability of the results based upon the individual replicates, we elected to do an additional analysis based upon all of the available data for the 50 minute blocks. For each of the blocks, the data from the four replicate samples were combined into a single data set, and the analyses were repeated. These analyses bring all of the available data to bear on the hypothesis of interest.

The top portion of Table 6 presents the results of fitting the models to the full data for the Grade 8 data. The pattern of results for the full data differs somewhat from that found in the main analysis. The difference test was significant for comparing to 50 minute Block to Block 25_A ($\chi^2(8) = 16.97$, $p < .05$). However, the comparison to Block 25_B was not significant ($\chi^2(8) = 9.17$, $p > .05$). Although there is

some evidence of misfit, there was not sufficient evidence to conclude that the 50 minute block has different relationships with the background variables. Based on these results, we would conclude that the 50 minute Block and the 25 minute blocks yielded similar relationships of reading with background variables at Grade 8, supporting our decision to include it in the operational analysis of the 1992 NAEP reading assessment.

Insert Table 6 about here

The middle portion of Table 6 presents the results of fitting the models to the full data for Block 50_A at Grade 12. The pattern of results was similar to that obtained for Grade 8. The difference test was significant for comparing to Block 50_A to Block 25_A ($\chi^2(8) = 19.33$, $p < .05$). However, the comparison to Block 25_B was not significant ($\chi^2(8) = 14.20$, $p > .05$). As was the case with Grade 8, there is not sufficient evidence to conclude that Block 50_A has different relationships with the background variables. Based on these results, we would conclude that Block 50_A and the 25 minute blocks yielded similar relationships of reading with background variables at Grade 12, supporting our decision to include it in the operational analysis of the 1992 NAEP reading assessment.

The bottom portion of Table 6 presents the results of fitting the models to the full data for Block 50_B at Grade 12. The difference test was not significant for comparing Block 50_B to Block 25_A ($\chi^2(8) = 10.26$, $p > .05$). Similarly, the comparison to Block 25_B was not significant ($\chi^2(8) = 10.81$, $p > .05$). Based on these results, we would conclude that Block 50_B and the 25 minute blocks yielded similar relationships of reading with background variables at Grade 12, supporting our decision to include it in the operational analysis of the 1992 NAEP reading assessment.

Discussion

At this point, it is appropriate to explore two limitations of the present study. The first limitation involves the nature of the data structure, and the research question at the heart of this study. The second limitation involves the reliance upon likelihood ratio-based χ^2 statistics.

A clear limitation of the present study is the mismatch between the research question of interest and the data structure. As we pointed out in the introduction, the data at hand are not ideal to answer the question, "Do the 25 minute and 50 minute reading blocks measure the same trait?" The optimal data collection design would involve presenting (with appropriate counter-balancing) 25 minute and 50 minute blocks to the same examinees. These data would allow factor analytic and IRT models to be fit, which could directly answer the question. A special study with this design would be ideal. However, practical issues, such as the need to limit assessment time to approximately one hour (to maintain the voluntary school participation in NAEP) and limited funds render such a special study infeasible.

Given the data at hand, the present study was undertaken in the spirit that some test of the assumptions underlying scaling is better than no test. We agree with Campbell and Stanley (1966, p. 35) that:

> *The task of theory-testing data collection is therefore predominantly one of rejecting inadequate hypotheses*. In executing this task, any arrangement of observations for which certain outcomes would disconfirm theory will be useful... . [emphasis in the original]

This work presents an opportunity for us to disconfirm the hypothesis that the two types of blocks measure the same construct. Having here failed to reject that hypothesis, we may proceed with a little more faith in the results of the operational analysis of the NAEP reading assessment than if we hadn't examined the hypothesis.

The reliance upon χ^2-based statistics is also a clear limitation in this study (although we have

attempted to circumvent some of the weaknesses though the use of replications and the sensitivity analyses). Both the ML and GLS statistics in structural equation modeling are derived based upon assumptions of multivariate normality. This assumption clearly does not hold for our data; three of the variables are dichotomous. Given the large differences found for the replicate analyses, one is left wondering to what extent the differences observed might be due to violations of the assumption of multivariate normality. Also, the χ^2-statistic ignores NAEP's complex sample, and so overstates the significance of the χ^2-difference test.

As discussed in Hu et al. (1992), recent statistical work has indicated the asymptotic appropriateness of the ML and GLS χ^2 statistics. However, the sample sizes required for asymptotic properties to hold is still unknown. Our choice of GLS was guided by the empirical findings in Hu et al. Furthermore, Muthén (personal communication, May 22, 1992) indicates that this failure of normality is probably not a serious limitation in the present context. Also, Bentler (1985) indicates that failures of normality tend to increase the χ^2 statistic. Empirical results by Donoghue, MacKinnon, Pentz, and Pentz (1987) support this contention. Therefore, one might infer that, in the present context, the failure of assumptions of multivariate normality should increase the confidence in the results which lead us to include the 50 minute blocks in the operational NAEP analysis, although see Hu et al. for a counter-argument.

However, a decision rule which is highly robust to violations of normality would be helpful. Hu et al. (1992) found that the scaling-corrected index of Sartorra and Bentler (1988a,b) worked well across the conditions they examined. The index is not available in the LISREL 7 program, however. Also, we do not know the applicability of the index as an index of incremental fit (e.g., χ^2-difference) for testing nested models; Hu et al. examined only a single model.

An obvious alternative to over-reliance on tests of the χ^2 statistic is the bootstrap (e.g., Efron, 1982; Stine, 1990). We are currently attempting to extend this study by applying the bootstrap

methodology to the present context. The application of the bootstrap to confirmatory factor analysis models is relatively complex (e.g., Bollen & Stine, 1993). Problems in implementation prevented the inclusion of that work in this paper, but the work is currently on-going.

References

Bollen, K. A., & Stine, R.A. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models*. Newbury Park, CA: Sage Publications.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally College Publishing Company.

Donoghue, J. R., MacKinnon, D. M., Pentz, C. A., & Pentz, M. A. (1987). Reliability of reported drug use under varying assumptions. Paper presented at the annual meeting of the Western Psychological Association, Long Beach, CA, April 1987.

Efron, B. (1982). *The bootstrap, the jackknife, and other resampling plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Hu, L-t., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351-362.

Joreskog, K. G. & Sorbom, D. (1988). *PRELIS: A preprocessor for LISREL (2nd ed.)*. Mooresville, IN: Scientific Software, Inc.

Joreskog, K. G. & Sorbom, D. (1989). *LISREL 7 user's reference guide*. Mooresville, IN: Scientific Software, Inc.

Messick, S., Beaton, A., & Lord, F. (1983) National Assessment of Educational Progress reconsidered: A new design for a new era. NAEP Report 83-1. Princeton, NJ: Educational Testing Service.

National Assessment Governing Board (1991). *Reading framework for the 1992 National Assessment of Educational Progress: NAEP Reading Consensus Project*. Washington, DC: National Assessment Governing Board, U. S. Department of Education.

Rao, J. N. K., & Thomas, D. R. (1991). Chi-squared tests with complex survey data subject to misclassification error. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (Eds.), *Measurement errors in surveys*, pp. 637-663. New York: John Wiley & Sons.

Sartorra, A. & Bentler, P. M. (1988a). Scaling corrections for chi-square statistics in covariance structure analysis. *American Statistical Association 1988 proceedings of the Business and Economics Section* (pp. 308-313). Alexandria, VA: American Statistical Association.

Sartorra, A. & Bentler, P. M. (1988b). Scaling corrections for statistics in covariance structure analysis. (UCLA Statistics Series 2). Los Angeles, CA: University of California, Department of Psychology.

Stine, R. (1990). An introduction to bootstrap methods: Examples and ideas. In J. Fox and J. S. Long (Eds.) *Modern methods of data analysis*. Newbury Park, CA: Sage Publications, pp. 325-373.

Table 1*

Numbers of Examinees and Items for Reading Blocks used in Main Analyses

Grade	Block	Number of Examinees	Number of Items	Multiple Choice	Short Constructed Response	Extended Constructed Response
8	25_A	575	14	7	6	1
	25_B	587	13	6	6	1
	50	582*	13	5	6	2
12	25_A	502	12	5	6	1
	25_B	494	10	3	6	1
	50_A	505*	16	10	4	2
	50_B	519*	12	7	2	3

* Sample of 25% of the examinees who took this block.

Table 2

Zero-order Correlations and Multiple R_2

For Predicting Block Total Score from Background Variables

	Grade 8			Grade 12			
Background Variable	25_A	25_B	50	25_A	25_B	50_A	50_B
Gender	.09	.07	.15	.08	.04	.08	.13
Black[1]	-.22	-.33	-.26	-.32	-.20	-.30	-.27
Hispanic[2]	-.28	-.22	-.20	-.14	-.13	-.10	-.09
Hours of TV	-.14	-.24	-.18	-.32	-.19	-.15	-.26
Parents' Education	.30	.28	.37	.31	.24	.28	.33
How Good a Reader Are You?	.32	.27	.30	.29	.23	.36	.30
How Many Items Did You Get Correct?	.41	.38	.40	.40	.40	.41	.40
How Hard Did You Try?	.17	.13	.09	.11	.11	.18	.13
Multiple R^2	.38	.35	.35	.35	.25	.37	.36

[1] This is an indicator variable, scored 1 if examinees identified themselves as Black/African American, and 0 otherwise.

[2] This is an indicator variable, scored 1 if examinees identified themselves as Hispanic, and 0 otherwise.

Table 3

Grade 8, 50 Minute Block Fit Statistics

		Corr. Free ($\chi^2(68)$)	Corr. Constrained ($\chi^2(76)$)	Difference ($\chi^2(8)$)
Test (GLS)	Comp. 25_A	111.82	122.45	10.63
	Comp. 25_B	98.44	110.81	12.37
	Null: 25_A v. 25_B	96.98	108.38	11.40
ML	Comp. 25_A	117.67	128.41	10.74
	Comp. 25_B	97.99	109.54	11.55
	Null: 25_A v. 25_B	101.06	112.25	11.19
ADF	Comp. 25_A	42.68	61.08	18.40*
	Comp. 25_B	33.90	52.39	18.49*
	Null: 25_A v. 25_B	37.39	60.21	22.82**
Impute Missing (GLS)	Comp. 25_A	96.88	105.92	9.04
	Comp. 25_B	85.36	98.68	13.32
	Null: 25_A v. 25_B	91.58	108.00	16.42*
Replication A	Comp. 25_A	99.54	115.60	16.06*
	Comp. 25_B	86.16	93.25	7.09
Replication B	Comp. 25_A	108.13	117.75	9.62
	Comp. 25_B	94.75	104.17	9.42
Replication C	Comp. 25_A	95.78	110.44	14.66
	Comp. 25_B	82.40	88.62	6.22

* $p < .05$
** $p < .01$

Table 4

Grade 12, Block 50_A Fit Statistics

		Corr. Free ($\chi^2(68)$)	Corr. Constrained ($\chi^2(76)$)	Difference ($\chi^2(8)$)
Test (GLS)	Comp. 25_A	123.32	138.08	14.76
	Comp. 25_B	133.00	148.74	15.76*
	Null: 25_A v. 25_B	152.71	160.63	7.92
ML	Comp. 25_A	125.51	139.97	14.46
	Comp. 25_B	142.64	155.60	12.96
	Null: 25_A v. 25_B	165.19	171.59	6.40
ADF	Comp. 25_A	53.27	74.64	21.37*
	Comp. 25_B	54.35	71.90	17.55*
	Null: 25_A v. 25_B	69.75	80.51	10.76
Impute Missing (GLS)	Comp. 25_A	93.22	115.15	21.93**
	Comp. 25_B	116.03	129.86	13.83
	Null: 25_A v. 25_B	118.99	130.24	11.25
Replication A	Comp. 25_A	120.26	146.11	25.85**
	Comp. 25_B	129.94	148.96	19.02*
Replication B	Comp. 25_A	142.99	164.60	21.61**
	Comp. 25_B	152.67	164.37	11.70
Replication C	Comp. 25_A	124.97	139.33	14.36
	Comp. 25_B	134.65	144.96	10.37

* $p < .05$
** $p < .01$

Table 5

Grade 12, Block 50_B Fit Statistics

		Corr. Free ($\chi^2(68)$)	Corr. Constrained ($\chi^2(76)$)	Difference ($\chi^2(8)$)
Test (GLS)	Comp. 25_A	138.62	145.86	7.34
	Comp. 25_B	148.31	160.09	11.78
	Null: 25_A v. 25_B	152.71	160.63	7.92
ML	Comp. 25_A	137.91	148.12	10.21
	Comp. 25_B	155.04	169.00	13.96
	Null: 25_A v. 25_B	165.19	171.59	6.40
ADF	Comp. 25_A	63.79	77.13	13.34
	Comp. 25_B	64.87	82.04	17.17*
	Null: 25_A v. 25_B	69.75	80.51	10.76
Impute Missing (GLS)	Comp. 25_A	107.44	117.32	9.88
	Comp. 25_B	130.25	142.20	11.95
	Null: 25_A v. 25_B	118.99	130.24	11.25
Replication A	Comp. 25_A	154.60	180.68	26.08**
	Comp. 25_B	164.28	191.28	27.00**
Replication B	Comp. 25_A	120.88	150.39	29.51**
	Comp. 25_B	130.56	152.53	21.97**
Replication C	Comp. 25_A	154.15	160.65	6.50
	Comp. 25_B	163.83	170.56	6.73

* $p < .05$
** $p < .01$

Table 6

Fit Statistics for Analyses Using Full Data

		Corr. Free ($\chi^2(68)$)	Corr. Constrained ($\chi^2(76)$)	Difference ($\chi^2(8)$)
Grade 8 (GLS)	Comp. 25_A	139.14	156.11	16.97*
	Comp. 25_B	125.76	134.93	9.17
Grade 12 Block 50_A (GLS)	Comp. 25_A	217.11	236.44	19.33*
	Comp. 25_B	226.80	241.00	14.20
Grade 12 Block 50_B (GLS)	Comp. 25_A	255.95	266.21	10.26
	Comp. 25_B	265.63	276.66	10.81

* $p < .05$
** $p < .01$